



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification⁶ :

G06F 17/27

A1

(11) International Publication Number:

WO 98/08169

(43) International Publication Date:

26 February 1998 (26.02.98)

(21) International Application Number: PCT/US97/14741

(81) Designated States: CN, JP, KP, KR, MN, SG.

(22) International Filing Date: 21 August 1997 (21.08.97)

Published

With international search report.

(30) Priority Data:

60/023,426

22 August 1996 (22.08.96)

US

(71) Applicant: INSO CORPORATION [US/US]; 31 St. James Avenue, Boston, MA 02116-4101 (US).

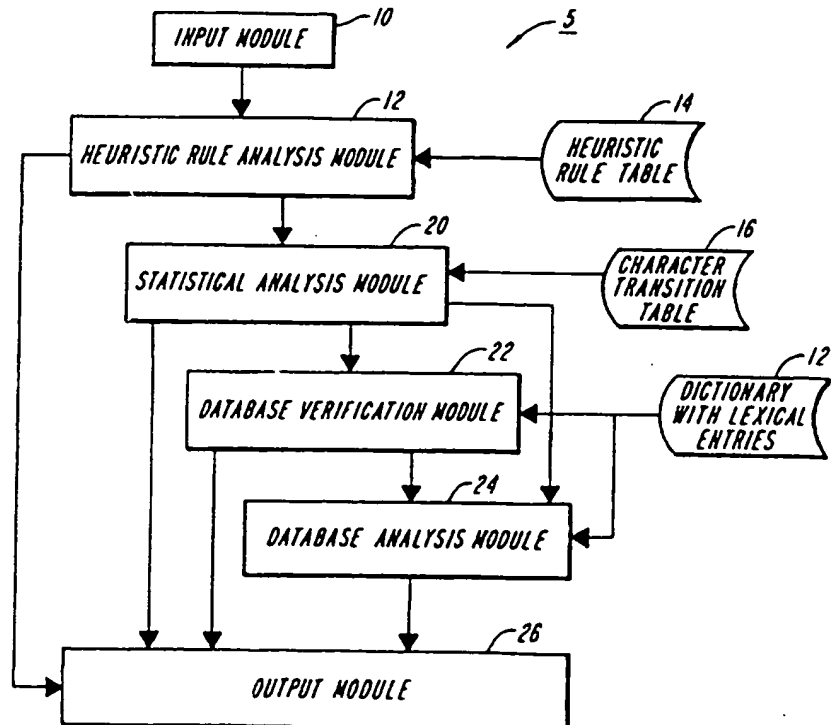
(72) Inventors: CARUS, Alwin, B.; 20 East Quinobequin Road, Newton, MA 02168 (US). WIESNER, Michael; 45 Maple Street, West Roxbury, MA 02132 (US). KRAUSE, Deborah; 48 Hill Circle, Burlington, MA 01803 (US).

(74) Agents: LIEPMANN, W., Hugo et al.; Lahive & Cockfield, LLP, 28 State Street, Boston, MA 02109 (US).

(54) Title: METHOD AND APPARATUS FOR BREAKING WORDS IN A STREAM OF TEXT

(57) Abstract

A word breaker utilizing a lexicon module and a processing module to identify word breaks in a stream of Asian (e.g. Japanese, Chinese, or Korean) language text. The lexicon module is a dictionary or database containing words native to the language of the input text. The processing module includes a plurality of analysis modules which operate on the input text. In particular, the processing module can include modules that analyze the input text using heuristic rules and statistical analysis to identify a first set of word breaks, thereby reducing the size of segments with undefined word breaks. The processing module also includes a database analysis module that identifies the remaining undefined word breaks in those smaller segments that have undergone heuristic or statistical analysis.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	IS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GII	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

METHOD AND APPARATUS FOR BREAKING WORDS IN A STREAM OF TEXT

5

FIELD OF THE INVENTION

The present invention relates to automated language analysis systems, and relates to such systems embodied in a computer for receiving digitally encoded text composed in a natural language. More particularly, the invention relates to an efficient and accurate method and apparatus for determining word breaks in digitally encoded Asian language text.

BACKGROUND OF THE INVENTION

Automated language analysis systems embedded in a computer typically include a lexicon module and a processing module. The lexicon module is a table of lexical information, such as a "dictionary" or database containing words native to the language of the input text. The processing module includes a plurality of analysis modules which operate upon the input text in order to process the text and generate a computer understandable semantic representation of the natural language text. Automated natural language analysis systems designed in this manner provide for an efficient language analyzer capable of achieving great benefits in performing tasks such as information retrieval.

Typically the processing of natural language text begins with the processing module fetching a continuous stream of electronic text from an input module. The processing module then decomposes the stream of natural language text into individual words, sentences, and messages. For instance, individual words in the English language can be identified by joining together a string of adjacent character codes between two consecutive occurrences of a white space code (i.e. a space, tab, or carriage return).

Japanese language text, and other Asian languages such as Chinese and Korean, can not be separated into individual words as easily as English language text. Asian language text typically includes a string of individual characters each separated by white-space. Words in these Asian languages are formed of a single character or a successive groups of characters, but the boundaries between the words are not explicitly identified in the written text. The written text does not clearly indicate whether any particular character forms a complete word or whether the particular character is only

part of a word. In addition, the written characters may be from one or more character alphabets. For example, Japanese words may be formed in one of three character types: Katakana, Hiragana, Kanji, and Romaji characters. Identifying these ambiguous word boundaries between the characters proves important in electronically translating or
5 processing Asian language documents.

Some prior art systems attempt to determine these word boundaries with simple pattern matching rules while other prior art systems resort to using a database of Asian language words to identify word breaks in Asian language text. For instance, US Patent No. 5,029,084, issued to Morohasi et al., discloses a system that combines various
10 pattern matching approaches to determine word boundaries in the text. The Morohasi system identifies character divisions based on character type definitions (i.e. Katakana, Kanji, Hiragana) and then processes the sentence by comparing the characters to a content word dictionary containing Japanese words. For any character segments remaining after this initial processing, a series of compound word synthesizing rules are
15 used to determine the division of the remaining segments. This system has the drawback of performing an up front costly comparison analysis of the characters in the stream of text with a content word dictionary of the Japanese language.

Other prior art systems use morpheme analysis to determine the word breaks in a Japanese language sentence. US Patent No. 5,268,840, issued to Chang et al., describes
20 a method and apparatus for morphologizing text. The Chang system discloses segmenting the input text of characters into the longest morphemes that can be formed from the input text. This is achieved by forming the longest morpheme from the remaining characters in the sentence which is listed in a dictionary of valid morphemes and determining if it is conjunctive with the previously divided morpheme. The
25 conjunctiveness of successive morphemes can be based upon grammar rules that require two adjacent morphemes to obey certain rules of connection.

Morphological analyzers of the type disclosed in Chang have efficiency problems. For example, subsequent identification of morphemes beyond the initially identified morpheme may indicate that the earlier identified morphemes are incorrect
30 and require further analysis. This inherent recursive nature of the system causes inefficiencies in the processing of the input text. In addition, the morphological analysis of Chang requires two separate processing steps. In the first step, the system identifies the morphemes themselves and in the second step the system requires the application of the morphological rules to the entire document. Thus, a morphological analysis system
35 typically requires considerable computer processing effort and frequent database

accessing resulting in longer processing times, coupled with the ever present risk of needing to review and reassess earlier faulty analysis.

Accordingly, an object of the invention is to provide a word breaker that efficiently and accurately identifies word breaks in a stream of Asian language text.

- 5 Other general and specific objects of the invention will be apparent and evident from the accompanying drawings and the following description.

SUMMARY OF THE INVENTION

- 10 The invention provides an apparatus and method for identifying word breaks in Asian language text that overcomes the limitations in the above described techniques. In particular, the present invention achieves accuracy and efficiency by applying computationally expensive procedures to word segments having unidentified word breaks only after less computationally expensive procedures have been exploited to
15 reduce the size of the word segments by unidentifying word breaks. The invention discloses an apparatus that first analyzes the stream of text with computationally inexpensive processing, thereby reducing the size of the segments with undefined word breaks within the stream of text. The inventive system then analyzes these smaller segments with more computationally expensive processing in a manner that does not
20 require reexamination of the earlier analysis steps.

- For instance, one aspect of the word breaker includes an element for storing the input character string, a first memory element for storing a character-transition table, a second memory element for storing a dictionary, a statistical analysis module for reducing the number of unidentified word breaks in the stream of text, and a database
25 analysis processor for locating the remaining unidentified words breaks in the stream of text. The statistical analysis module, a less computationally expensive process than a dictionary look-up, examines a first segment in the stream of text to locate a first word break. The statistical analysis module, using computationally inexpensive processes, then partitions the stream of text into at least a first sub-segment and a second sub-
30 segment divided by the first word break. The first and second sub-segments are then analyzed using the more computationally expensive database analysis processor to identify the remaining word breaks in the first segment. The database analysis processor identifies the remaining word breaks by comparing the characters in the first and second sub-segments with entries in the dictionary stored in the second memory element.

- 35 The invention further provides for a statistical analysis processor that identifies the first word break with the aid of data stored in the character-transition table. For

instance, the statistical processor can identify the first word break as a function of a statistical morpheme stored in the character-transition table. In one instance, the statistical processor can compare the characters in the first segment of the input stream with the entries in the character-transition table, and the statistical processor can then
5 identify those characters in the stream of text that match entries in the character-transition table. Once a match is found, the statistical processor can then associate a character-transition tag with the matched characters in the stream of text. The associated character-transition tag identifies the existence of a concatenation between successive characters, a break between successive characters, or an unknown transition between
10 successive characters.

Additional features of the statistical analysis processor provide for a windowing module for forming a window of successive characters from the stream of text. The windowing module also includes structure for comparing the window of successive characters with entries in the character-transition table. This allows a window of
15 characters to be compared with entries in the character-transition table and to identify whether the window of characters forms a statistical morpheme as registered in the character-transition table.

Further in accordance with the windowing module, the invention can further include a processing element for sliding the window of successive characters along the first segment of the stream of text. This provides for a system that can compare
20 successive groupings of characters within the stream of text with the entries in the character-transition table. Additionally, the windowing module can include a processing element that controls the number of characters in the window, that is the size of the sliding window. This allows various numbers of successive characters to be compared
25 with the entries in the character-transition table.

Another aspect of the invention includes a heuristic rule table and a heuristic rule module for further reducing the number of possible character combinations forming words in the input stream of text. The heuristic rule module and its associated heuristic rule table are typically computationally inexpensive procedures that are applied before
30 the database analysis processor is utilized, thereby reducing the size of word segments having unidentified word breaks that must be processed by the database analysis processor. The heuristic rule table and heuristic rule module can identify word breaks in the input character string; for Japanese, these are based upon numbers, punctuation, Roman letters, classifiers, particles, honorific prefixes, emperor years, and Kanji-
35 Katakana character-transitions.

Other features of the invention include a word verification module for verifying matches between identified words in the input character string, and a character-transition table that includes character strings of morphemes that form a minimum spanning set necessary to identify character-transitions in the input character string. Additionally, the
5 entries in the character-transition table can be of variable length.

One preferred method of the invention for locating unidentified word breaks in a input character string includes the steps of: storing the input character string, identifying a morpheme in a first segment of the input character string, reducing the number of unidentified word breaks in the input character string based upon the identified
10 morpheme, and locating the remaining unidentified words breaks in the input character string by comparing the segments of the input string with a dictionary. This method employs computationally expensive procedures only after less expensive computational procedures have been employed to reduce the size of the segments having unidentified word breaks. For instance, the inventive method first identifies a word break in the
15 input character string based upon a morpheme, the identified word break divides the input string into first and second sub-segments having the remaining unidentified word breaks. The first and second sub-segments are then analyzed using the computationally more expensive process of comparing the sub-segments with a dictionary.

An additional aspect of the invention provides for a machine readable storage
20 medium having instructions stored thereon that direct a programmable computer to implement the word breaker disclosed herein. The instructions on the machine readable storage medium implement a first element for reducing the number of unidentified word breaks in a character string by locating a first word break in a first segment of the character string as a function of at least one statistical morpheme in the first segment, the
25 first word break dividing the first segment into a first sub-segment and a second sub-segment, and a second element for locating substantially all of the remaining unidentified word breaks in the first and second sub-segments by comparing the first and second sub-segments with entries in a dictionary of lexical entries.

BRIEF DESCRIPTION OF THE DRAWINGS

- FIG. 1 is a block diagram of a word breaker according to the present invention;
- 5 FIG. 2 is a flow chart illustrating the processing steps of the heuristic rule analysis module of FIG. 1;
- FIG. 3 is a flow chart illustrating the processing steps of the statistical analysis module of FIG. 1;
- FIG. 4 is a detailed flowchart showing the steps of the database verification
- 10 analysis of FIG. 1;
- FIG. 5 is a flowchart illustrating the steps of the database analysis of FIG. 1; and
- FIG. 6 is a detailed flowchart illustrating the comparison step of the database analysis procedure of FIG. 5.

15

DETAILED DESCRIPTION

- FIGURE 1 illustrates a word breaker 5 in accordance with the invention.
- 20 The word breaker 5 comprises a processing module operably coupled with a lexicon module. The processing module can include a statistical analysis module 20, a database analysis module 24, an optional heuristic rule analysis module 12, and an optional database verification module 22. The separate sub-modules 12, 20, 22 and 24 of the processing module are operably coupled together to allow the transfer of data and
- 25 control signals between the sub-modules. The lexicon module can include a character-transition table 16, a dictionary 18 with lexical entries, and an optional heuristic rule table 14. The character-transition table 16 is operably coupled with the statistical analysis module 20, the dictionary 18 is operably coupled with the database verification module 22 and with the database analysis module 24, and the heuristic rule table 14 is
- 30 operably coupled with the heuristic rule analysis module 12. As further illustrated in FIG. 1, the word breaker 5 can include an input module 10 for receiving the stream of input text and an output module 26 for generating an output signal representative of the stream of input text with identified word breaks.

- The word breaker 5 can be implemented using electronic hardware devices,
- 35 software, or a combination of the two. For example, a digital computer running a UNIX platform can be loaded with software instructions to implement the structure and processes of the word breaker 5. The input module 10 can be a keyboard, a text

processor, a machine readable storage device, or any other structure capable of generating or transferring a stream of text. The processing module and its sub-modules 12, 20, 22 and 24 can be implemented through software instructions executed by the processor in the digital computer or by specifically designed hardware components implementing the equivalent instruction set. The lexical module and its sub-modules 14, 16 and 18 can be implemented by tables stored in a volatile or a non-volatile memory device that is operably coupled with the processing module. The output module can be implemented using any device capable of storing, displaying, or transferring the signals generated by the processing module of the invention. For instance the output module can include a SCSI hard drive, electronic memory devices, or a video monitor.

In operation, the input module 10 either receives or generates the input stream of text that requires identification of word boundaries. The input module 10 can either pre-process the text or it can directly transfer the stream of incoming text to the heuristic rule analysis module 12. For instance, the stream of Japanese text entering the input module 10 can be represented electronically using JIS (Japanese Industry Standard), Shift-JIS, EUC (Extended Unix Code) or Unicode, wherein each Japanese character is represented by two bytes, but the input data can contain single-byte or double-byte characters. It is preferable to have the Japanese text represented in one standard format, such as Unicode. Accordingly, the input module 10 advantageously provides pre-processing to convert the input stream of text into a standard format, such as Unicode.

The input module 10 can also associate tags with each character in the input stream of text. The tags associated with each electronic equivalent of the characters identify attributes of the characters. In particular, a character-transition tag can be associated with each character such that the character-transition tag identifies word breaks between characters in the input stream of text. For instance, in the present invention, character-transitions can be represented as a tri-state value with a 0 representing an unknown transition, a 1 representing an a break between the characters, and a 2 representing that there is a link, i.e., a concatenation, between the characters. Each transition is represented by a two-bit flag and accordingly the bit values for UNKNOWN, BREAK, and LINK are 00, 01, and 10 respectively. For an "N" character pattern, there are "N+1" of these bit pairs which are necessary to completely represent the transition. For example:

the pattern |A_B?C|;

wherein: " | " indicates a break,

" _ " indicates a link, and

" ? " indication an unknown transition;

can have character-transition tags represented as: 01 10 00 01.

The optional heuristic rule module 12 receives the stream of text from the input module 10 and transmits a heuristically processed stream of text to the statistical analysis module 20. Alternatively, the heuristic module can be removed from the word breaker 5 such that the stream of text output by the input module 10 passes directly to the statistical analysis module 20.

In general, the heuristic rule analysis module 12 identifies a character-transition in the character stream such that the number of possible character combinations forming words in the character string are reduced. For example, the heuristic rule module can identify character-transitions based on a classification of the character type, wherein the character-transitions identify either word breaks between successive characters, a link between successive characters, or an unknown transition between successive characters. The heuristic rule module 12 acts in cooperation with the heuristic rule table 14 to identify the character-transitions in the character stream.

Once the character-transitions are identified the heuristic rule module sets the character-transition tags associated with each identified character-transition to either a break between characters, or a link between characters. The heuristic rule module 12 then passes all character-transition tag data to the statistical rule module 14. The statistical rule module reads the character-transition tag data and analyzes only those segments having unknown interior character-transitions.

The statistical analysis module 20 either receives a stream of data from the optional heuristic rule analysis module 12 or from the input module 10. The statistical analysis module reduces the number of unidentified word breaks in the character string based upon a statistical morpheme in the character string. A statistical morpheme is a morpheme that is identified based upon a statistical analysis of the frequency of occurrence of the morpheme in a corpus of text. The statistical analysis module cooperates with the character-transition table 16 to identify the statistical morphemes in the input character stream.

In operation, the statistical analysis module 20 analyzes all remaining unknown character-transitions in the input text by comparing a sequence of characters containing unknown character-transitions in the input text to sequences of characters stored in the character-transition table 16. The character-transition table 16 contains entries of character segments, i.e. statistical morphemes, which are chosen to statistically predict a character-transition based on the order of characters. Accordingly, a character-transition tag associated with the input character stream can be set when a sequence of characters

in the input text match an entry in the character-transition table. In those aspects of the invention wherein the word breaker 5 utilizes both the heuristic analysis module 12 and statistical analysis module 20, it is estimated that approximately 90% of the character-transitions in the input character stream are accurately identified.

5 After processing of the stream of text by the statistical analysis module 20, the input character stream can be either processed by the database verification module 22, the database analysis module 24, or directly by the output module 26. For instance, when the input character stream has no remaining unidentified character-transitions after processing by the statistical analysis module 20, processing proceeds directly to the
10 output module 26.

 The database verification module 22 is an optional module that compares the input text segments identified as words by the character breaks, to lexical entries contained in the dictionary 18. If the word is correct and properly verified it is passed to the output module. After processing by the verification module, which includes
15 adjustments or corrections of spurious word forms that remain after statistical analysis, approximately 95% of the character-transitions are accurately identified. The remaining unknown input text segments with character-transitions are passed to the database analysis module 24. The database analysis module 24 compares the remaining unknown character-transitions to the lexical entries in the dictionary 18 to determine the final
20 word breaks. The text, now with spaces between the words, is passed to the output module 26.

 FIG. 2 illustrates a flow diagram for implementing the heuristic rule analysis module 12. The heuristic rule module in cooperation with a heuristic rule table identifies characters in the text that consistently identify character-transitions, i.e., a break between
25 characters or a link between characters. For example, in one embodiment of the present invention the heuristic rule module identifies character-transitions formed by numbers, Roman letters, punctuation marks, classifiers, particle delimiters, repeat characters, honorific prefixes, Kana/Kanji combinations, and end of sentence markers.

 After a character-transition is identified, the heuristic rule module sets a
30 character-transition tag associated with each character to either a break or a link. Once a character-transition tag is set, the character-transition tag need not be modified by any further processing because of the accuracy of the processing used to set the character-transition tags. The results from the heuristic rule analysis module 12 may be one of three types: 1) an identified word; 2) a segment with unknown interior character-
35 transitions; 3) and an unidentified segment. An identified word is one in which there is a break before the first character and after the last character with all character-transitions

set as a link in the interior of the identified word. A segment with unknown character-transitions will have a break before the first character and after the last character, but will have at least one unknown character-transition in the interior of the segment. An unidentified segment is a succession of characters having an insufficient number of character-transitions identified to qualify as either an identified word or a segment with unknown interior character-transitions.

With further reference to FIG. 2, the heuristic rule module begins execution at step 29 wherein the input data stream is transferred to the heuristic rule module 12 for processing. After step 29, step 30 is executed.

Step 30 analyses the input stream of text with a set of heuristic rules for identifying character-transitions in the stream of text based on the classification of the characters. In one aspect of the invention, this is accomplished by comparing the characters in the input stream of text with a classification system or set of rules stored in either the heuristic rule table 14 or the heuristic rule analysis module 12. For example, in one embodiment the classification system classifies the characters according to whether the characters consist of numbers, Roman letters, punctuation marks, classifiers (i.e., currency characters, time and date characters), particle delimiters, end of sentence characters, Kanji and Kana repeat characters, katakana/kanji combination characters, or emperor year characters. Based on the classification of any particular character, a character-transition tag associated with the particular character can be set, as shown in Step 32 of FIG. 2.

All forms of numbers--single and double-byte Arabic numerals, kanji standard and bank digits, and power of ten characters --are identified as a word segment by the heuristic rule module 12. In addition, numbers followed by any of the time, emperor year, or currency classifiers would be marked as a number and time, date, or currency respectively.

The heuristic rule module 12 also identifies a string of single-byte Romaji letters (whether English, Western European, Greek, or Cyrillic) as one word segment. The segment begins at the first Romaji letter the word breaker encounters and ends at the last consecutive Romaji letter.

The heuristic rule module also places word breaks before and after each character in the stream of text that is classified as a punctuation mark. Although, commas and decimal points used in numbers will not have a word break placed before or after the comma or decimal point. Table 1 is a representative table that identifies whether characters in the stream of text are to be classified as punctuation marks.

Table 1.

Punctuation Marks

Phrasal Marks

[d = double-byte; s = single-byte]

・	(Japanese comma)	:	(s)
。	(maru)	:	(d)
、	(d)	:	(s)
。	(s)	?	(d)
、	(d)	?	(s)
。	(s)	!	(d)
、	(d)	!	(s)
:	(d)	...	(d)

Range marks:

～	(d)
—	(d)
・	(d)
-	(s)
—	(d-byte minus sign)
/	(d)
\	(d)

Quotation marks

“	(d)	[(d)
”	(d)]	(d)
・	(d)	”	(d)
・	(d)	”	(d)
『	(d)	”	(s)
』	(d)	”	(s)

Brackets & Parentheses

{	»
:	{
€	}

Table 1. (cont.)

<			
>			
[)	(d)
]		([s]
[(s))	[s]
]	(s)	([d]
[(d))	[d]
		((d)
)	(d)

Other symbols/punctuation marks

°	(d --- degree)	¶	(d)
°C	(d --- degree centigrade)	¶	(d)
%	(d)	¶	(d)
%	(s)	¶	(d)
¶	(d)	¶	(d)
¶	(s)	¶	(d)
¶	(d)	¶	(d)
¶	(s)	¶	(d)
¶	(d)	¶	(d --- minutes)
¶	(s)	¶	(d --- seconds)
¶	(d)	¶	(d --- zip code symbol)
¶	(s)	¶	(d)
¶	(d)	¶	(s)
¶	(s)	¶	(d)
¶	(d)	¶	(s)
¶	(s)	¶	(d)
¶	(d)	¶	(d)
¶	(d)	¶	(number, phrase repeat sign)
¶	(d)	¶	(symbol which replaces "しめ" or "しめ")

Bullets

★	○	◆	□	▽
☆	●	◇	△	▼
※	◎	■	▲	§

Table 2 represents a list of classifiers, both general and specific classifiers (such as time and currency) used by the heuristic rule module 12 to identify character-transitions. In general a classifier will follow the number expression it is classifying. Only those classifiers identified in Table 2 as "currency classifiers Part 2" precede the number expression. Any time the number is preceded or followed by a classifier the number expression includes the classifier as part of the word. That is, there should be no break between the number and the classifier.

Table 2.

Classifiers

The asterisk in the tables below indicates that the classifier can also be a word by itself

General Classifiers

便*	45b1	川*	5d4e	図所	7bS7.G240
言*	53f0	羽*	7f0d	ヶ所	30fG.G240
段*	65b5	きれ*	304d.30S3	図所	300b.G240
度*	5ea6	品*	7f3G	ヶ所	30fS.G240
取*	673f	バック*	30d1.30c3.30af	カ所	30ab.G240
花*	62S3	かん*	304b.3093	切れ	5207.30S3
箱*	7bb1	カン*	30ab.30f3	ヶ月	30fG.6705
登*	707a	版*	74fG	ヵ月	30fS.6705
香*	73G3	ピン*	30d3.30f3	ヵ月	30ab.6705
水*	672c	びん*	3073.3093	図月	7bS7.6705
国*	5Gde	時間*	6G42.9393	図糸	7bS7.67G1
野*	Sed2	年度*	5c74.5ea6	ヶ糸	30fG.67G1
脚*	S11a	様*	7a2e	ヵ糸	30fS.67G1
公*	540d	匹	5339	カ糸	30ab.67G1
両*	4e21	匹	7bS7	ヶ国	30fG.552d
位*	7c92	個	500b	ヵ国	30fS.5Gfd
山*	5c71	枚	679a	カ国	30ab.5Gfd
注*	793e	枚	6d41	カ国	30ab.5Gfd.8a9e
組*	7d44	枚	6b73	ヶ国	30fG.5Gfd.8a9e
足*	Sdb3	局	518a	ヵ国	30fS.5Gfd.8a9e
機*	G3c3	袋	9Gbb	図	8a9e
描い*	G3c3.3044	周	54GS	ニ	3053
袋*	SSSb	機	8256	ニ	30b3
芯*	7740	機	633a	図	92ad
丁*	4e01	通	901a	セント	30bb.30f3.30cS
頭*	9S2d	ヶ	30fG	図	3323
機*	GSdf	ヶ	30b1	図	5272
人*	4eba	つ	30G4		
機*	G43f	機	9320		
登*	7573	才	624d		
機*	770e	階	9GSe		
所*	5c02	件	4cfc		

* 糸 67G1 has been removed as of 7/29/96

Table 2. (cont.)

Time/Date Classifiers: (These always FOLLOW numbers)

時	6642	Hour
時半	6642.9593	Half Hour
分	5206	Minute
日	65E5	Day
月	6705	Month
年	5E74	Year
秒	79D2	Second

Currency Classifiers Part 1: (These FOLLOW the number)

円	5186	Yen
ドル	5f17	Dollar
ドル	30c9.30eb	Dollar
フラン	3326	Dollar
マルク	30d5.30e9.30f3	Franc
ペソ	30de.30eb.30af	Mark
元	30da.30bd	Peso
ウォン	5143	Yuan
元	30a6.30aa.30f3	Won
ルーブル	5713	Yuan
リラ	30eb.30fc.30d6.30eb	Ruble
??(not supported by Shift-JIS)	30ea.30e9	Lira
??(not supported by Shift-JIS)	3306	Won (Korea)
??(not supported by Shift-JIS)	3307	Escudo (Portugal)
??(not supported by Shift-JIS)	3313	Guilders
??(not supported by Shift-JIS)	331a	New Cruzeiro (Brazil)
??(not supported by Shift-JIS)	331b	Krone (Generic)
??(not supported by Shift-JIS)	331d	Koruna (Czech Republic)
??(not supported by Shift-JIS)	3321	Schilling (Austria)
??(not supported by Shift-JIS)	3340	Pound (Great Britain, etc.)
??(not supported by Shift-JIS)	3350	Yuan (People's Republic of China)
??(not supported by Shift-JIS)	3352	Lira (Italy, San Marino, Vatican)
??(not supported by Shift-JIS)	3353	Rupee (India)
??(not supported by Shift-JIS)	3354	Squared Ruble
??(not supported by Shift-JIS)	3335	Squared Franc
??(not supported by Shift-JIS)	3337	Squared Peso
??(not supported by Shift-JIS)	3336	Squared Penih
??(not supported by Shift-JIS)	333a	Squared Pence
??(not supported by Shift-JIS)	3346	Squared Mark

Table 2. (cont.)

Currency Classifiers Part 2: (These PRECEDE number)

S	0024	USD
¥	00A5	JPY
¥	FFE5	JPY
S	FF04	USD
£	FFE1	PND
¥	FE69	JPY same as single-byte YEN 00A5
??	FFE6	KPW (Korean Won sign - not available in shift-JIS)

- Table 3 represents a list of particle delimiters used by the heuristic rule analysis module 12 to identify further character-transitions in the stream of text. The right hand column lists rules specific to the particle or function shown in the left hand column. The
- 5 rules listed in the right hand column of Table 3 are translated according to the following code: BB = break immediately before the particle; BA = break immediately after the particle; K[n] = kanji; T = katakana; H = hiragana; and Pnc = punctuation mark.

Table 3. Particle/Function Word Delimiters

3.1. Particle/Function Word Delimiters for Word-breaking

Delimiter [pronunciation]	Unicode Value	Rule
を [wo]	3092	BB BA
ヲ [wo]	30f2	BB BA
迄 [made]	5fc4	BB BA
然り [shirari]	7136.306a	BB BA if followed by a K
尚 [nao]	5c1a	BB BA if followed by a comma
だが [daga]	3060.304c	BB BA only if at BOS (preceded by maru)
なあ [naa]	306a.3041	BB BA
例えば [tatoeba]	4f5b.3048.3070	BB BA
然るに [shikaruni]	7136.306b.306b	BB BA
すなはち [sunawachi]	3059.306a.306f.3061	BB BA
なぜならば [nazenaraba]	306a.305c.306a.3059.3070	BB BA
それゆえに [soreyueni]	305d.305c.3056.3045.306b	BB BA
それどころか [soredokoruka]	305d.305c.3069.3053.305d.304b	BB BA

Table 3. (cont.)

舌しくは (moshikuwa)	62e5.3057.304f.306f	BB BA
乃至 (naishi)	4e43.51f3	BB BA
或は (aruwa)	6216.306f	BB BA
即ち (sunawanchi)	5373.3061	BB BA
但し (tadashi)	4f46.3057	BB BA if not followed by "空"
或いは (aruwa)	6216.3044.306f	BB BA
あるいは (aruwa)	3042.305b.3044.306f	BB BA
それでは (soredewa)	305d.305c.3057.306f	BB BA
それなら (sorenara)	305d.305c.305a.3059	BB BA
それとも (soretomo)	305d.305c.3055.3052	BB BA
すなわち (sunawanchi)	3059.306a.305f.3061	BB BA
が (ga)	304c	<p>BB BA if preceded by K1 && followed by K2 && "K1 or K2" is NOT in gaword.tbl</p> <p>BB BA if preceded by K && followed by T or Pnc && K is NOT any of { 0x6211 # 𐤁𐤍𐤏𐤃 0x543e # 𐤁𐤍𐤏𐤃 0x6b6f # 𐤁𐤍𐤏𐤃 0x4e57 # 𐤁𐤍𐤏𐤃 }</p> <p>BB BA if preceded by T && followed by K or T or Pnc</p> <p>BB BA if preceded by Pnc && followed by K or T</p> <p>BA if preceded by H && followed by T</p> <p>BA if preceded by H && followed by K && K is NOT in gasuffix.tbl</p> <p>BB if preceded by T && followed by H</p>
は (wa)	306f	<p>BB BA if preceded && followed by K</p> <p>BB BA if preceded && followed by T</p> <p>BB BA if preceded by T && followed by K or Pnc</p> <p>BB BA if preceded by L && followed by T or Pnc && K is NOT any of { 0x62a6 # 𐤁𐤍𐤏𐤃 0x53c5 # 𐤁𐤍𐤏𐤃 }</p>

Table 3. (cont.)

		0x5b9f #実) BA if followed by K && K is NOT (0x666b #虫) BA if followed by T BB if preceded by T BB if preceded by K && K is NOT any of { 0x62a6 #返 0x53c8 #又 0x5b9f #実)
へ [e]	3078	BB BA if preceded by K or T or Pnc && followed by K or T or Pnc BA if preceded by H && followed by T BB if preceded by T && followed by H
の [no]	306c	BB BA if preceded && followed by K && "K の K" is NOT in noword.tbl BB BA if preceded by K && followed by T or Pnc BB BA if preceded by T && followed by K or Pnc or T BA if preceded by H && followed by T BB if preceded by T && followed by H
まで [made]	307e, 3067	BB BA if preceded by T or K or Pnc BA if preceded by H && followed by K or T or Pnc
にて [nite]	306b, 3066	BB BA if preceded by K && followed by K or T or Pnc BB BA if preceded by T && followed by K or T or Pnc
だけ [dake]	3060, 3051	BB BA if preceded by K && followed by K or T BB BA if preceded by T && followed by K or T or Pnc
こそ [koso]	3053, 305d	BB BA if preceded by Pnc && followed by K or T
すら [sura]	3059	BB BA if preceded by K or T or Pnc && followed by K or T or Pnc
や [ya]	3064	BB BA if preceded by K1 or T or Pnc && followed by K2 or T or Pnc && K1 is NOT (0x51b7 #何)
なり [nari]	306a, 305a	BB BA if preceded && followed by K
で	3067	BB BA if preceded && by followed by T BB BA if preceded by K1 && followed by K2 or Pnc or T && K1 is NOT any of { 0x6c5a, #居 0x79c0, #了 0x51fa, #出 0x594f, #奏 0x6c56, #戻

Table 3. (cont.)

	0x982d, #頭
	0x64ab, #眉
	0x786c, #碌
	0x8339, #面 }
&& K2 is NOT any of	
	{ 0x6e6f, #湯
	0x7acb, #立
	0x5207, #切
	0x7121, #無
	0x7389, #玉
	0x7db2, #綳
	0x53e3, #口
	0x5c71, #山
	0x77f3, #石
	0x76e4, #呂
	0x5408, #合
	0x9eba, #夏 }
BB BA if preceded && followed by T	
BB BA if preceded by T and followed by K or Pnc	
BB if preceded by T	
BB if preceded by K which is NOT any of {	0x6c5a, #汚
	0x79c0, #秀
	0x51fa, #出
	0x594f, #奏
	0x6c55, #泥
	0x982d, #頭
	0x64ab, #眉
	0x786c, #碌
	0x8339, #面 }
BA if followed by T	
BA if followed by K which is NOT any of {	0x6e6f, #湯
	0x7acb, #立
	0x5207, #切
	0x7121, #無
	0x7389, #玉
	0x7db2, #綳
	0x53e3, #口
	0x5c71, #山
	0x77f3, #石
	0x76e4, #呂
	0x5408, #合
	0x9eba, #夏 }

Table 3. (cont.)

と [to]	3068	<p>BB BA if preceded by T or Pnc && followed by K or T or Pnc</p> <p>BB BA if preceded by K1 && followed by K2 or T or Pnc && K1 is NOT in toprefix.tbl && K2 is NOT any of { 0x8ffd #追 0x71c3 #煙 0x611b #愛 0x77f3 #石 0x5ea7 #座 0x9803 #頂 0x76ae #皮 }</p> <p>BB if preceded by T</p> <p>BB if preceded by K && K is NOT in toprefix.tbl</p> <p>BA if followed by T</p> <p>BA if followed by K && K is NOT any of { 0x8ffd #追 0x71c3 #煙 0x611b #愛 0x77f3 #石 0x5ea7 #座 0x9803 #頂 0x76ae #皮 }</p>
から [kara]	304b.3089	<p>BB BA if preceded by K or T or Pnc && followed by T or Pnc</p> <p>BB BA if preceded by T or Pnc && followed by K && K is NOT in karasuf.tbl</p> <p>BB BA if preceded by K1 && followed by K2 && K1 is NOT any of { 0x4ed8 #付 0x5132 #能 0x5206 #分 0x52a9 #助 0x61f8 #能 0x639b #掛 0x67b6 #架 0x61ea #自 0x9060 #遠 0x9810 #預 } && K2 NOT in karasuf.tbl</p> <p>BB if preceded by T</p> <p>BB if preceded by K && K is NOT any of { 0x4ed8 #付 0x5132 #能 0x5206 #分 0x52a9 #助 0x61f8 #能 0x639b #掛 0x67b6 #架</p>

Table 3. (cont.)

		0x61ea #目 0x9060 #透 0x9S10 #預) BA if followed by T BA if followed by K && K is NOT in karasuf.tbl
ニ [ni]	306b	BA if followed by T BA if followed by K && K is NOT any of { 0x76ee, #目 0x80a1, #透 0x76ae, #預 }
も [mo]	30S2	BB BA if preceded by T && followed by T or K or Pnc BB BA if preceded by Pnc && followed by T or K BB BA if preceded by K1 && followed by K2 && K1 is NOT in mcprefix.tbl && K2 is NOT any of { 0xS19c #透 0xG39b #預 } BB if preceded by K && K is NOT in mcprefix.tbl BA if followed by K && K is NOT any of { 0xS19c #透 0xG39b #預 } BA if followed by T BB BA if followed by Pnc
です		

Table 4 represents a list of Kanji and Kana (i.e. Kana is equivalent to Katakana and Hiragana) repeat characters used by the heuristic rule analysis module 12 to identify additional character-transitions in the stream of text. None of the repeat characters identified in Table 4 may start a word.

5

Table 4.

Repeat Characters

↵ = unc 3005; sjs S158; jis 2139; euc a1b9
↘ = unc 309d; sjs S154; jis 2135; euc a1b5
↙ = unc 309e; sjs S155; jis 2136; euc a1b6
↖ = unc 30fd; sjs S152; jis 2133; euc a1b3
↗ = unc 30fe; sjs S153; jis 2134; euc a1b4

Table 5 represents the honorific prefixes utilized by the heuristic rule analysis module 12 to identify additional character-transitions in the stream of text. If character [i] is an honorific prefix look at [i+1] and [i-1] and see if that character is found in the following table. If found, there is no break between any character, if only one character 5 is found place a break between the character that was found and the honorific prefix.

Table 5.
Honorific Prefix

Suffix Table

歌	6b4c	仁	4ec1	所	6240	幸	5e78	座	5ea7
伽	4f3d	腦	51b3	前	524d	形	5f62	面	8597
中	4e2d	用	7325	浜	6d5c	光	5149	宅	5b65
意	610f	陸	9675	船	8239	活	6d12	寢	89a7
衣	5563	科	6399	庭	8535	亡	4ea1	疎	7834
宇	5b57	談	8a60	証	8358	大	5927	寺	8005
苑	82d1	弊	5f0a	國	5712	邑	5cb3	号	5f13
度	6176	法	6cd5	器	5d69	宿	5bbf	虎	864e
史	53f2	坊	574a	津	6d25	曹	66f9	姓	5f37
史	74bd	段	6bbf	辻	6756	の	306e	候	4fa0
聖	55fd	三	4e09	路	8abf	任	8ec2	祖	7956
		女	98df	息	606f	麻	7c3e		
		修	4fec	統	7d71	候	994c		

Prefix Table

久	4e45	姉	59c9	身	8cab
副	5236	姐	59d0	運	9054
北	5317	崩	5d29	防	9632
大	5927	櫻	6e21		
女	5973	統	7d71		

A character switch between Kanji to Katakana or Katakana to Kanji usually indicates a word break with some exceptions. Table 6 represents the exceptions to this rule. For this rule to apply there must be a string of at least 4 or more katakana characters. When there are less than 4 katakana characters in a string, the heuristic rule module 12 does not place any breaks.

If the Kanji character immediately preceding a katakana string of 4 or more characters matches with any of those in table 6, the heuristic rule module 12 does not place any breaks, otherwise the heuristic rule module 12 places a break between the kanji and the first katakana character. If the kanji immediately following a katakana string of 4 or more characters matches with any of those in Table 7, the heuristic rule module 12 does not place any breaks, otherwise the heuristic rule module 12 places a break between the kanji and the last katakana

Table 6.
Katakana/Kanji Combination

Exception Prefix

不 4e0d	併 4ff3	四 56db	昨 6628
丹 4e39	再 518d	六 5927	未 6765
登 4e7e	北 5317	小 5c0f	京 6771
亞 4e9c	南 5357	麻 5ec3	次 6b21
今 4ecc	双 53cc	急 6025	造 6b6f
低 4f4e	台 542b	急 6025	造 6bb5
濟 6e80	紙 7d19	超 5d85	非 975e
炭 70ad	脫 8131	逆 9006	高 9ad8
熟 71b1	螺 6776	過 904e	鼻 9f3b
省 7701	被 88ab	全 91d1	
穴 7a74	低 8910	銀 9260	
等 7b49	西 897f	銀 92f3	

Table 7.

Exception Suffix

人	4eba	带	5e2f	新	6e0S	茶	8336
党	515a	座	5ca7	海	6e7e	庄	8349
司	52G4	座	5eab	共	72bG	西	83cc
化	5316	奔	5f01	用	752S	也	866b
器	56GS	式	5f0f	病	75c5	製	86fd
型	578b	形	5fG2	庭	75c7	計	8a0S
基	57fa	性	G027	的	76S4	話	8a9e
空	5S02	教	G559	目	76ee	說	8a2c
場	5S34	展	G591	石	77f3	及	Scde
恒	5SG9	族	G5cf	法	793e	發	8cea
家	5bbG	星	G61f	科	79d1	發	917S
局	5c5e	技	Ga5f	秒	79d2	發	925b
山	5c71	油	Gcb9	管	7ba1	發	9271
岩	5cn9	法	6cd5	管	7cd6	鋼	92fc
昌	5cfG	派	Gd3e	紙	7d19	紙	9G4d
川	5ddd	流	Gd41	票	7d20	預	9S5c
州	5dde	海	Gd77	題	8171	區	9ce5
市	5e02	設	Gdb2	理	S266		

Table 8 represents the characters delimiting an end of sentence that are utilized by the heuristic rule module 12 to identify character-transitions in the stream of text.

Table 8. End of Sentence

The end of sentence is marked by the following characters:

- 。 -- maru is the primary EOS marker.
- 。J -- maru followed by a close quotation mark can indicate EOS, except when it is followed by a particle.
- 。 -- a close quotation mark followed by white spaces can indicate EOS when no maru can be found.
- (space) -- Since Japanese does not use spacing between words, a space indicates a break in a clause and it is most likely to be the end of sentence in the absence of maru. This is often seen at the end of a paragraph.

Accordingly, under step 30 of FIG. 2, the heuristic rule analysis module 12 searches for character-transitions enumerated in the above tables and identifies the character-transitions as either a break between characters or a link between characters. A break between characters will be also be a break between words. A link between characters will indicate a potential internal connection between characters within a character segment. Under step 32, the heuristic module 12 sets the character-transition tags associated with characters in the stream of text based upon the analysis performed under Step 30. Thereafter, at Step 34, the heuristic rule module 12 stores the set or identified determined character-transition tag associated with a particular character segment.

At Step 35, the heuristic rule module 12 passes all character-transition tag data to the statistical rule module 14. The statistical rule module will read the character-transition tag data, and only analyze those segment with unknown interior character-transitions. Subsequent processing of the character string after the processing by heuristic rule module 12 does not include processing of those character-transitions identified by the statistical rule analysis module 12.

FIG. 3 illustrates a flow chart for implementing the statistical analysis module 20. The flow chart details a process whereby a window of successive characters in the stream of text is compared with the entries in a character-transition table. If the statistical analysis module identifies a match between the window of characters and an entry in the character-transition table, the statistical analysis module sets those character-transition tags associated with the characters in the current window. If no matches are found in the current window of characters, then the characters can be subjected to further analysis by the word breaker 5. After the current window of characters is processed, the statistical analysis module 20 slides the window across the input stream of text to review a different grouping of characters.

As shown in FIG. 3, sliding window analysis begins at Step 40 wherein the statistical rule module 20 receives a segment of character having at least one undefined character-transition therein. After Step 40, control proceeds to Step 60 for initialization. In step 60, the character-transition tag before the first character in the segment and after the last character in the segment are set to a break. This allows the system to approach the entire character group as an individual word. As processing continues, the default character-transition tags associated with the characters in the segment may be overwritten to identify either links or word breaks. Once a character-transition is changed from its default value to those values identifying either links or word breaks, the character-transition can not be changed. After Step 60, control proceeds to Step 62.

At Steps 62 and 64 the size of the sliding window is initialized. In the preferred embodiment, the sliding window has a maximum size of four characters and is decremented by one through each loop. Accordingly, at Step 62 a counter "i" is initialized to the starting address of the input text, and a counter "j" is set to one less than the desired window length. At Step 64, the window length is defined to run from "i" to "i+j".

At Step 66, the characters in the sliding window are compared with entries in the character-transition table 16 to identify character-transitions associated with the characters in the segment.

The character-transition table 16 is derived from analysis of a hand-broken corpus of approximately 200,000 character segments. The table 16 stores reliable information regarding high-frequency character-transitions. The corpus can include unigram, digram, trigram and quadgrams patterns. In the present invention, the corpus containing the data for the character-transition table has the following contents: character sequence; break pattern; frequency of break pattern; connection pattern; and frequency of connection pattern.

The corpus containing the data for the character-transition table is first run through the heuristic rule analysis module 12 to remove any segments which can be analyzed by the existing heuristic rules. This prevents the statistical analysis module 20 from duplicating the efforts of the heuristic analysis module 12.

Initial testing of the character-transition table data using the full hand-broken corpus showed the following results:

	<i>all patterns</i>	<i>ambiguous patterns</i>	<i>unambiguous patterns</i>	<i>percentage of unambiguous patterns</i>
<i>unigrams</i>	2,389	1,632	757	32%
<i>digrams</i>	57,069	9,732	47,337	83%
<i>trigrams</i>	157,116	7,316	149,800	95%
<i>quadgrams</i>	222,443	3,614	218,829	98%

The character-transition table in the present invention preferably comprises data on unigram [single character-transitions], digram [two character-transitions], trigram [three character-transitions], and quadgram [four character-transitions], which leads to a maximum of ten (10) bits of information attached to the quadgram pattern.

In the character-transition table a n-gram pattern would be represented by "n" linked tables. In the present invention a 4 table mechanism would be needed. In the present invention, the first table of the character-transition table 14 would be:

Unicode-character-code start-pos number.

The first field represents the unicode value of the character code. start-pos represent the portion at which the rest of the pattern starts in the next (i.e., 2nd) table, and number is the number of patterns which have this character as the starting character. Therefore the number of n-gram patterns which start with the character in the first field and having sub patterns in the second table is represented by the value of the number field. The structure of the rest of the tables is similar except that a uni-gram pattern would be having a /0 in the second table and the 2nd field would have the bit map of the pattern. Similarly, the digram patterns would have the /0 in the first field of the third table and the bit map in the 2nd field of the third table.

With continued reference to FIG. 3, after the comparison process with the character-transition table at Step 66, control proceeds to Decision Step 68. If no match with an entry in the character-transition table is found, then logical flow proceeds to Step 70. If a match with an entry in the character-transition table is found, then logical flow proceeds to Step 78.

At Step 70, the window length variable is reduced by one thereby reducing the size of the window of characters. At Step 72, the size of the window is reviewed to determine whether the window still exists. If the window includes at least one character, then processing flows back to Step 64 through Action Box 76. The result of Steps 70, 72, and 76 is to incrementally reduce the size of the sliding window in an attempt to match the characters in the window with smaller entries in the character-transition table. However, if the window is reduced to the point where it does not include any characters and a match in the character-transition table has not been found, then processing flow to Steps 74 and 83.

At Step 74, the counter "i" is incremented by 1 and the counter "j" is set to 3. This causes the sliding window of successive character to slide to the right by one, and resets the size of the sliding window to four characters in length. After Step 74, control returns to Action Box 76.

At Step 83, those segments in the character string that still have undefined character-transitions after being processed by the Statistical Analysis module 20, are passed to the Database analysis module 24 for further processing.

At Step 78, the statistical analysis module 20 stores the new character-transition tag values for the matched character segments. After Step 78, the statistical analysis module 20 increments the memory location counter "i" and resets the window length variable to three. at Step 80. After Step 80, control proceeds to Step 82. At Step 82,

the first and last character-transitions in the next segment having an undefined character-transition is set to a break. After Step 82, control returns to Step 64.

Example of Heuristic Analysis and Sliding Window Statistical Analysis

5

The input stream of text 12NEWBOOKS needs to be analyzed. The heuristic rule module 12 identifies the number 12 and then passes the segment NEWBOOKS on to the statistical analysis module 20.

10

The Character-transition table 16, for this example, contains the following character-transitions:

[N E W] 01111101

[W B] 010001

[B O O] 00111101

15

[K S] 111101

Step 1:

The initial bitmap representation for NEWBOOKS is set to 0001010101010100 because only the first and last transition are known as breaks at this point:

20

00N01E01W01B01O01O01K01S00

First Window:

Step 2:

25

The counter i is initialized to 1.

Step 3:

The sliding window is defined as [1 2 3 4] = [N E W B].

Step 4:

30

The sequences for the current window which can be looked up in the Character-transition Table are:

[N E W B] [N E W] [N E] [N]

(1) [N E W B] is looked up in the Character-transition Table and not found.

5 (2) [N E W] is looked up in the Character-transition Table; it is found with the bitmap 01111101 which is merged with the existing bitmap:

00N01E01W01B01O01O01K01S00 +
01N11E11W01 =
10 00N11E11W01B01O01O01K01S00

Second Window:

Step 5:

i is set to 2.

Step 3:

15 The sliding window is defined as [2 3 4 5] = [E W B O].

Step 4:

The sequences for the current window which can be looked up in the Character-transition Table are:

20 [E W B O] [E W B] [E W] [E]

(1) [E W B O], [E W B], [E W], and [E] are looked up in the Character-transition Table and not found.

25 **Third Window:**

Step 5:

i is set to 3.

Step 3:

The sliding window is defined as [3 4 5 6] = [W B O O].

30 **Step 4:**

The sequences for the current window which can be looked up in the Character-transition Table are:

35 [W B O O] [W B O] [W B] [W]

(1) [W B O O] and [W B O] are looked up in the Character-transition Table and not found.

(2) [W B] is found with the bitmap 010001 which is merged with the existing bitmap:

40

00N11E11W01B01O01O01K01S00 +
01W00B01 =
00N11E11W00B01O01O01K01S00

Fourth Window:**Step 5:***i* is set to 4.**Step 3:**

5 The sliding window is defined as [4 5 6 7] = [B O O K].

Step 4:

The sequences for the current window which can be looked up in the Character-transition Table are:

10 [B O O K] [B O O] [B O] [B]

(1) [B O O K] is looked up in the Character-transition Table and not found.

(2) [B O O] is found with the bitmap 00111101 which is merged with the existing bitmap:

15

$$\begin{array}{rcl} 00N11E11W00B01O01O01K01S00 & + & \\ & & 00B11O11O01 \\ 00N11E11W00B11O11O01K01S00 & = & \end{array}$$

20

Fifth Window:**Step 5:***i* is set to 5.**Step 3:**

The sliding window is defined as [5 6 7 8] = [O O K S].

25

Step 4:

The sequences for the current window which can be looked up in the Character-transition Table are:

30 [O O K S] [O O K] [O O] [O]

(1) [O O K S], [O O K], [O O], and [O] are looked up in the Character-transition Table and not found.

Sixth Window:

35

Step 5:*i* is set to 6.**Step 3:**

The sliding window is defined as [6 7 8] = [O K S].

Step 4:

40

The sequences for the current window which can be looked up in the Character-transition Table are:

[O K S] [O K] [O]

(1) [O K S], [O K], and [O] are looked up in the Character-transition Table and not found.

5

Seventh Window:

Step 5:

i is set to 7.

Step 3:

10 The sliding window is defined as [7 8] = [O K S].

Step 4:

The sequences for the current window which can be looked up in the Character-transition Table are:

15 [K S] [K] .

(1) [K S] is found with the bitmap 111101 which is merged with the existing bitmap:

20
$$\begin{array}{r} 00N11E11W00B11O11O01K01S00 \\ \quad \quad \quad 11K11S01 \\ \hline 00N11E11W00B11O11O11K11S00 \end{array}$$

All transitions in the segment bitmap are now either '00' or '11', i.e., **breaks** or **connections**. The identified segments are thus:

25

$00N11E11W00B11O11O11K11S00 \Rightarrow | N E W | B O O K S |$

30 The statistical rule analysis module 20 splits the input segment NEWBOOKS into two segments NEW and BOOKS and identifies all segment-internal transitions. This means that the two new segments do not have to be analyzed anymore, and the segments are passed directly to the output module 26.

35 The sliding window output may produce any one of three different results for an input segment. The first, all transitions are identified as either a break or a link, and each entire segment may be verified by the database verification module 22 with a single lookup. The second, some transitions are identified. In this case, any identified transitions will be passed onto the database analysis module 24, which reduces the overall processing required to identify the word breaks. The third instance would be where no character-transitions are identified. In this instance, a full analysis by the database analysis module 24 is required.

FIG. 4 illustrates a flow chart for implementing the database verification module 22. An input is supplied from the statistical analysis module 20, which is an identified work, i.e., a character segment with a break before the first character, a break after the last character, and all internal character-transitions are defined as a link between characters. In step 90, the character segment is compared with lexical entries in the dictionary 18. In step 92 if a match occurs, the character segment is sent to the output module 26. If no match occurs the character string is passed to the database analysis module 24 for a full database analysis. It will be appreciated by those skilled in the art, that since the input text segment has been identified as a word, the comparing step, step 90, will compare the input text segment to one or two entries from the dictionary 18. Hence, since the number of lexical entries used is so small, any known method of sorting may be used to locate the lexical entries used to verify the identified word.

FIG. 5 illustrates steps for implementing the database analysis module 24. At step 83 the database analysis module receives a segment of characters having unidentified character-transitions. The database analysis module can receive the segment from either the statistical analysis module 20 or the database verification module 22. In step 100, the character segment is compared with lexical entries in the dictionary 18 to identify a match for the character segment in the dictionary. Once a match for the character segment is identified, the unidentified character-transitions in the word segment are identified and the completed segment is forwarded to output module 26 with all word breaks being identified.

FIG. 6 illustrates a flow chart detailing the comparison Step 100 of FIG. 5. At Step 120, the character segments with unknown character-transitions are input to the Database Analysis Module 24. After Step 120, control proceeds to Step 122.

At Step 122, a dictionary-look-up string is created based upon the input data received at Step 120. Afterwards, at Step 124, all left sub strings of the dictionary look up string are chosen for analysis. After Step 124, control proceeds to Step 126.

At Step 126, all words found in the dictionary 18 that possibly match the dictionary look up string are stored in a candidate word list. Then, at Step 128, the candidate word list is sorted in order of priority. The database analysis module 24 can sort the candidate word list based on various criteria, such as the starting position in the candidate word list or the length of the entries stored in the candidate word list. After Step 128, flow proceeds to Step 130.

At Step 130, the first entry in the sorted candidate word list is compared to the dictionary look up string. At decision box 132, the database analyzer 24 determines whether there is a match. If a match is found, then control proceeds to Step 138 where

the final word boundaries are set and all character-transition tags are set to either a break or a link, as determined by the matched entry in the candidate word list. The next dictionary look-up-string is then retrieved and the process repeated.

- If a match is not found at Decision Box 132, then control proceeds to Decision
- 5 Box 134. At decision box 134, the processor branches control to Step 136 if there are more words in the candidate word list, and the processor branches to Step 140 if there are no more words in the candidate word list. At Step 136, the next word from the candidate word list is retrieved for comparison with the dictionary lookup string. At
- 10 Step 140 the next substring is retrieved from the dictionary look up string for the comparison at Step 130.

While the invention has been shown and described having reference to specific preferred embodiments, those skilled in the art will understand that variations in form and detail may be made without departing from the spirit and scope of the invention. Having described the invention, what is claimed as new and secured by letters patent is:

CLAIMS

1. A method for locating unidentified breaks between words in an input character string formed of a plurality of characters, the method comprising the successive steps of
5 storing said input character string in a computer memory element,
identifying at least one morpheme in a first segment of said stored character string,
reducing the number of unidentified word breaks in said stored character string by locating a first word break in said first segment of said stored character string based
10 upon said at least one morpheme, said first word break dividing said first segment into a first sub-segment and a second sub-segment, and
locating further unidentified word breaks in said first and second sub-segments by comparing said first and second sub-segments to entries in a dictionary.
- 15 2. The method of claim 1, wherein said reducing step further includes verifying said first word break by matching a word preceding said first word break with a first entry in said dictionary and by matching a word following said first word break with a second entry in said dictionary.
- 20 3. The method of claim 1, wherein said identifying step includes the steps of locating word breaks and character-transitions by applying a set of rules to said stored character string to identify said at least one morpheme.
- 25 4. The method of claim 3, wherein said applying step further comprises forming a window of successive characters from said stored character string, comparing said window of successive characters to entries in a character-transition table, and
identifying said window of successive characters that matches an entry in the character-transition table as said at least one morpheme.
- 30 5. The method of claim 4, further comprising the step of decreasing the size of said window of characters if no entries in said character-transition table match said window of successive characters.

6. The method of claim 4, further comprising the step of sliding the window of successive characters across said stored character string if no entries in said character-transition table match said window of successive characters.
- 5 7. The method of claim 4, including the step of forming the character-transition table by generating a minimum spanning set of character strings necessary to identify character-transitions.
8. The method of claim 7, wherein the spanning set of character strings includes a
10 plurality of character strings having different lengths.
9. The method of claim 1, wherein said reducing step includes the steps of detecting a first character-transition in said stored character string based upon said at least one morpheme, and
15 locating said first word break as a function of said at least one morpheme and said first character-transition.
10. The method of claim 9, wherein said locating step includes the step of concatenating a first character and a second character together when said first character-transition indicates the existence of a connection between characters.
20
11. The method of claim 9, wherein said locating step further comprises the step of identifying a break between a first character and a second character when said first character-transition indicates the existence of a break between characters.
25
12. The method of claim 1, wherein said locating step further comprises the steps of creating a lookup string from characters within said first sub-segment, identifying a dictionary entry that substantially matches said lookup string, and
30 marking a second word break between the matched lookup string and a character that precedes the lookup string and marking a third word break between the matched lookup string and a character that follows the lookup string.
13. A method according to claim 12, further comprising the steps of creating a
35 candidate word list from a dictionary as a function of said lookup string, and wherein

said identifying step includes comparing an entry in said candidate word list with said lookup string.

14. The method of claim 12, further comprising the step of
5 validating that the matched lookup string is a word.

15. The method of claim 14, wherein the step of validating the matched lookup string includes
selecting an identified word, from the matched lookup string, and
10 comparing said matched lookup string to a dictionary for determining the validity of the identified word .

16. The method of claim 1, further comprising the step, prior to said identifying step, of applying a set of heuristic rules to said stored character string to identify a character-
15 transition in said first segment of said stored character string, said identification of a character-transition reducing the number of possible character combinations forming words in said stored character string.

17. The method of claim 16 further comprising the step of identifying a
20 concatenation between characters in said first segment as a function of said heuristic rules.

18. The method of claim 16 further comprising the step of selecting said heuristic rules for identifying a break between characters in said first segment.
25

19. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
locating a number in said stored character string, and
identifying a character-transition that precedes and a character-transition that
30 follows said located number.

20. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
locating identifying punctuation in said stored character string, and
35 identifying a character-transition that precedes and a character-transition that follows said located punctuation.

21. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
locating identifying Roman letters in said stored character string, and
5 identifying a character-transition that precedes and a character-transition that follows said located Roman letters.
22. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
10 locating identifying classifiers in said stored character string; and
identifying a character-transition that precedes and a character-transition that follows said located classifiers.
23. The method of claim 16, wherein said step of applying the set of heuristic rules
15 further comprises
locating identifying particles in said stored character string, and
identifying a character-transition that precedes and a character-transition that follows said located particles.
- 20 24. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
locating identifying honorific prefixes in said stored character string, and
identifying a character-transition that precedes and a character-transition that follows said located honorific prefixes.
- 25 25. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
locating an identifying emperor year in said stored character string, and
identifying a character-transition that precedes and a character-transition that
30 follows said located emperor year.
26. The method of claim 16, wherein said step of applying the set of heuristic rules further comprises
locating identifying Kanji-Katakana character-transitions in said stored character
35 string, and

identifying a character-transition that occurs at said located Kanji-Katakana character-transition.

27. A programmable computer an apparatus for locating unidentified breaks between
5 words in an input character string, comprising
- A) a computer memory element for storing the input character string,
 - B) first memory means for storing a character-transition table including character segments of morphemes,
 - C) second memory means for storing a dictionary, said dictionary including
10 lexical entries,
 - D) a statistical analysis module operably coupled with said first memory means storing character-transition table for reducing the number of unidentified word breaks by locating a first word break in a first segment of said input character string as a function of at least one statistical morpheme in said first segment, said first word break
15 dividing said first segment into a first sub-segment and a second sub-segment, and
 - E) a database analysis module operably coupled with said dictionary for locating substantially all of the remaining unidentified word breaks in said first and second sub-segments by comparing said first and second sub-segments with entries in said dictionary.
- 20
28. The apparatus of claim 27, wherein said statistical analysis module further comprises
- first processing means for identifying said at least one statistical morpheme in said first segment by comparing said first segment with entries in said character-
25 transition table and for detecting a character-transition associated with said at least one statistical morpheme, and
 - second processing means for locating a first word break in said first segment as a function of said at least one statistical morpheme and said character-transition.
- 30
29. The apparatus of claim 28, wherein said first processing means further comprises a windowing module for forming a window of successive characters from said first segment such that said window of characters can be compared with entries in said character-transition table.
- 35

30. The apparatus of claim 29, wherein said first processor module includes means for sliding said window of successive characters along said first segment of said input character string.

5 31. The apparatus of claim 29, further comprising means for changing the size of said window of characters.

32. The apparatus of claim 28, further comprising means for associating a character-transition tag with characters in said input string.

10

33. The apparatus of claim 32, wherein said means for associating a character-transition tag includes means for indicating a concatenation between successive characters.

15 34. The apparatus of claim 32, wherein said character-transition tag indicates a break between successive characters.

35. The apparatus of claim 27, wherein said database analysis module further comprises:

20 third processing means for identifying a match between said first sub-segment and an entry in said dictionary, and

fourth processing means for locating a second word break in said first sub-segment as a function of said matched entry.

25 36. The apparatus of claim 27, further comprising:
a heuristic rule table including a set of heuristic rules,
a heuristic rule module operably coupled with said heuristic rule table for identifying a character-transition in said first segment of said stored character string, such that the number of possible character combinations forming words in said stored
30 character string are reduced.

37. The apparatus of claim 27, further comprising a word verification module, operably coupled with said dictionary, for verifying matches between an identified word in said input character string and dictionary entries.

35

38. The apparatus of claim 27, wherein said character-transition table includes character strings of morphemes that form a minimum spanning set necessary to identify character-transitions in said input character string.

5 39. The apparatus of claim 38, wherein the spanning set includes a plurality of character strings having different lengths.

40. A machine readable data storage medium, comprising
means for reducing the number of unidentified word breaks in a character string
10 by locating a first word break in a first segment of said character string as a function of
at least one statistical morpheme in said first segment, said first word break dividing said
first segment into a first sub-segment and a second sub-segment, and
means for locating substantially all of the remaining unidentified word breaks in
said first and second sub-segments by comparing said first and second sub-segments
15 with entries in a dictionary of lexical entries.

41. The machine readable data storage medium of claim 40, further comprising a character-transition table including character segments of morphemes.

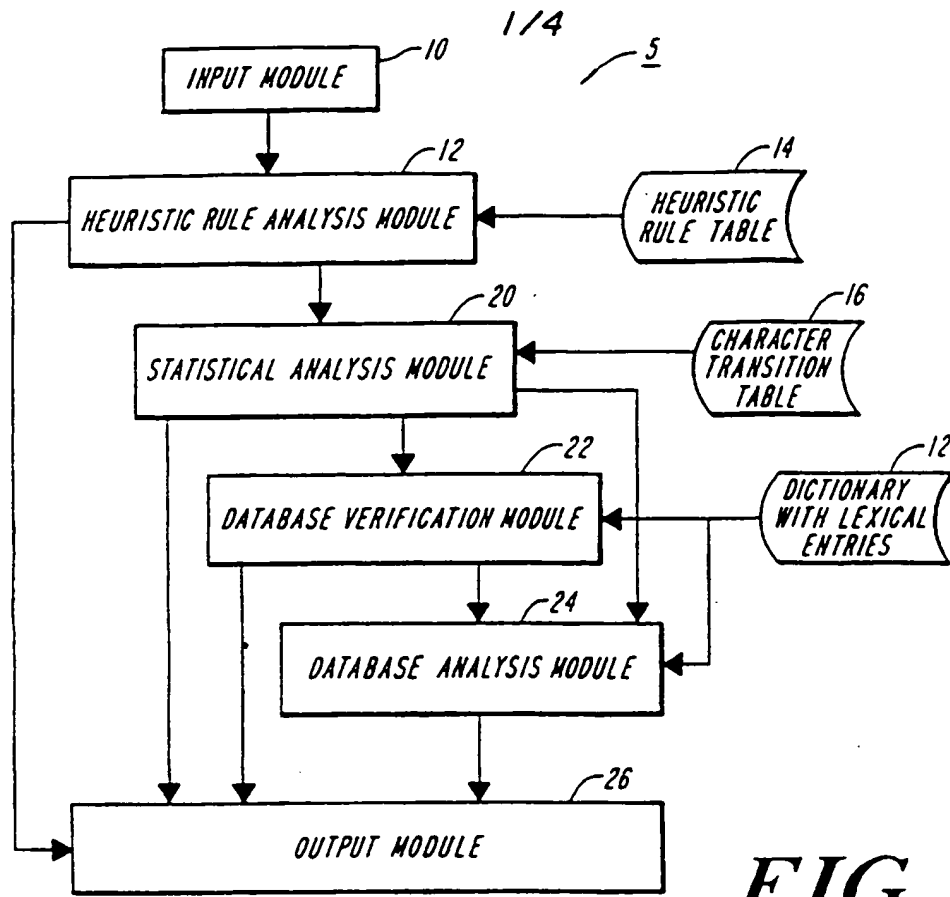


FIG. 1

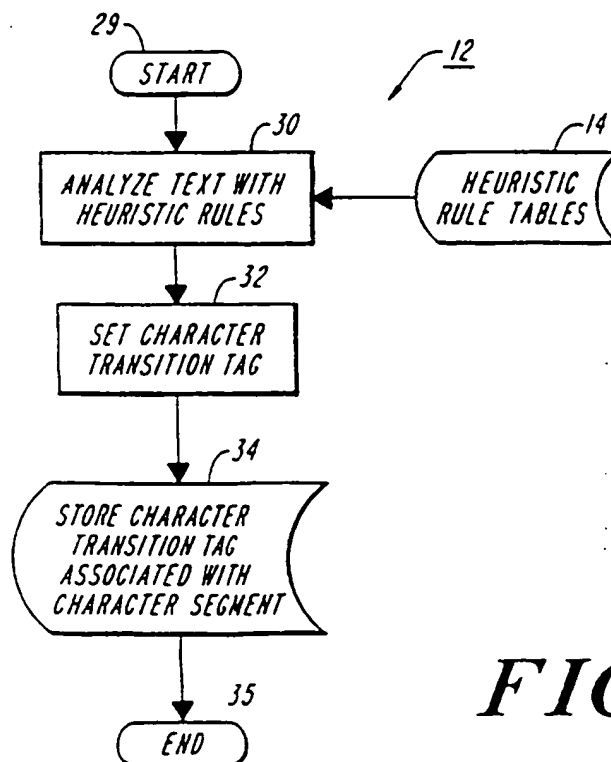
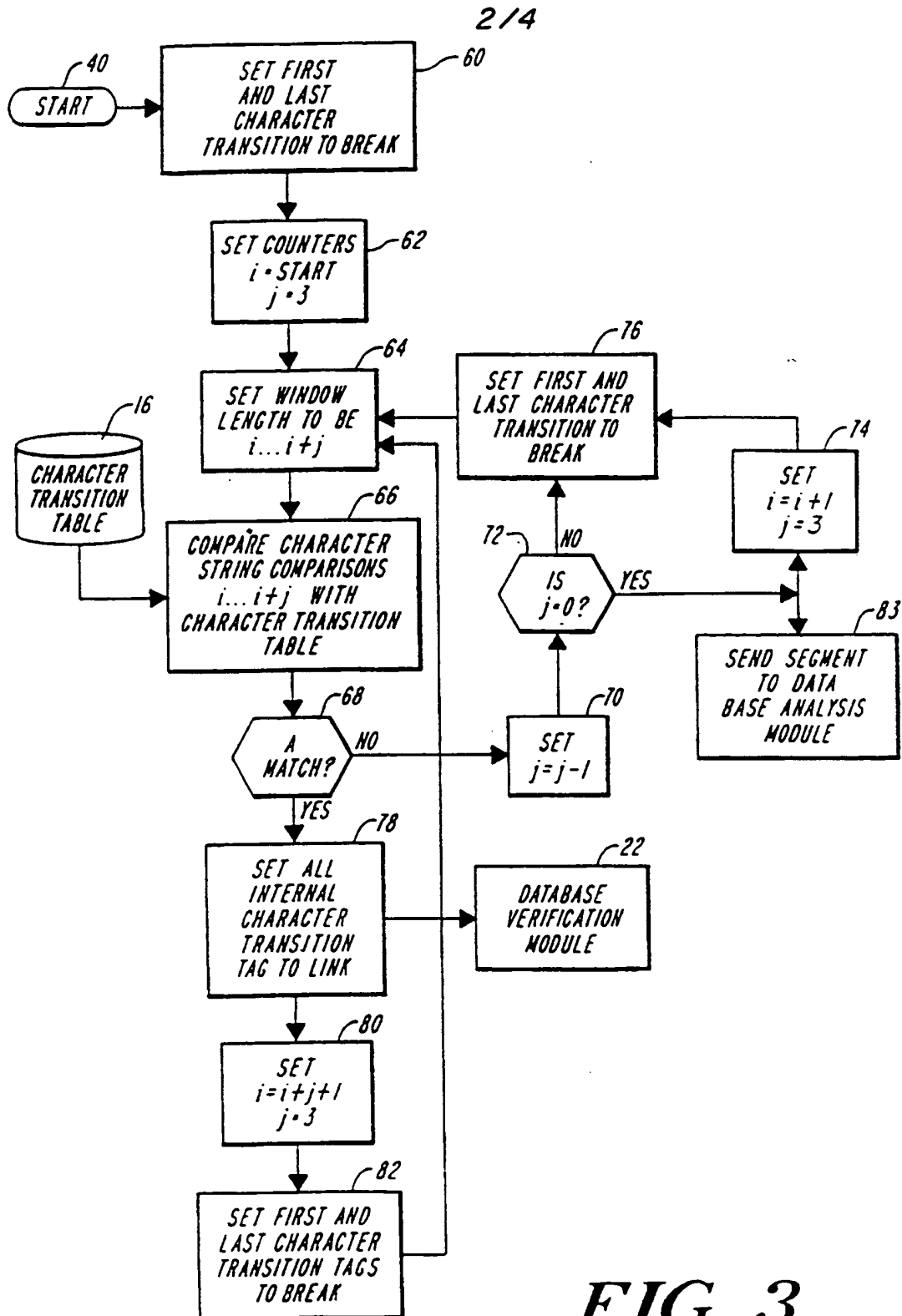
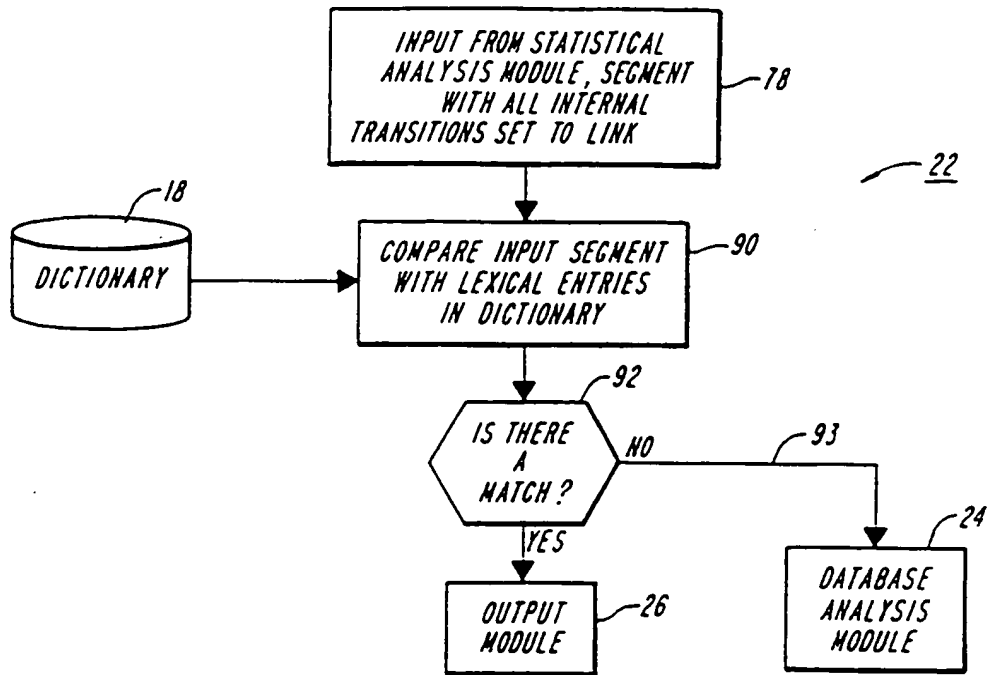
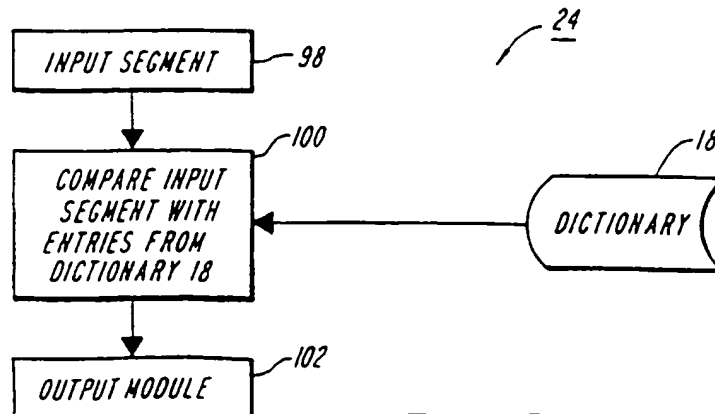


FIG. 2

**FIG. 3**

3/4

**FIG. 4****FIG. 5**

4/4

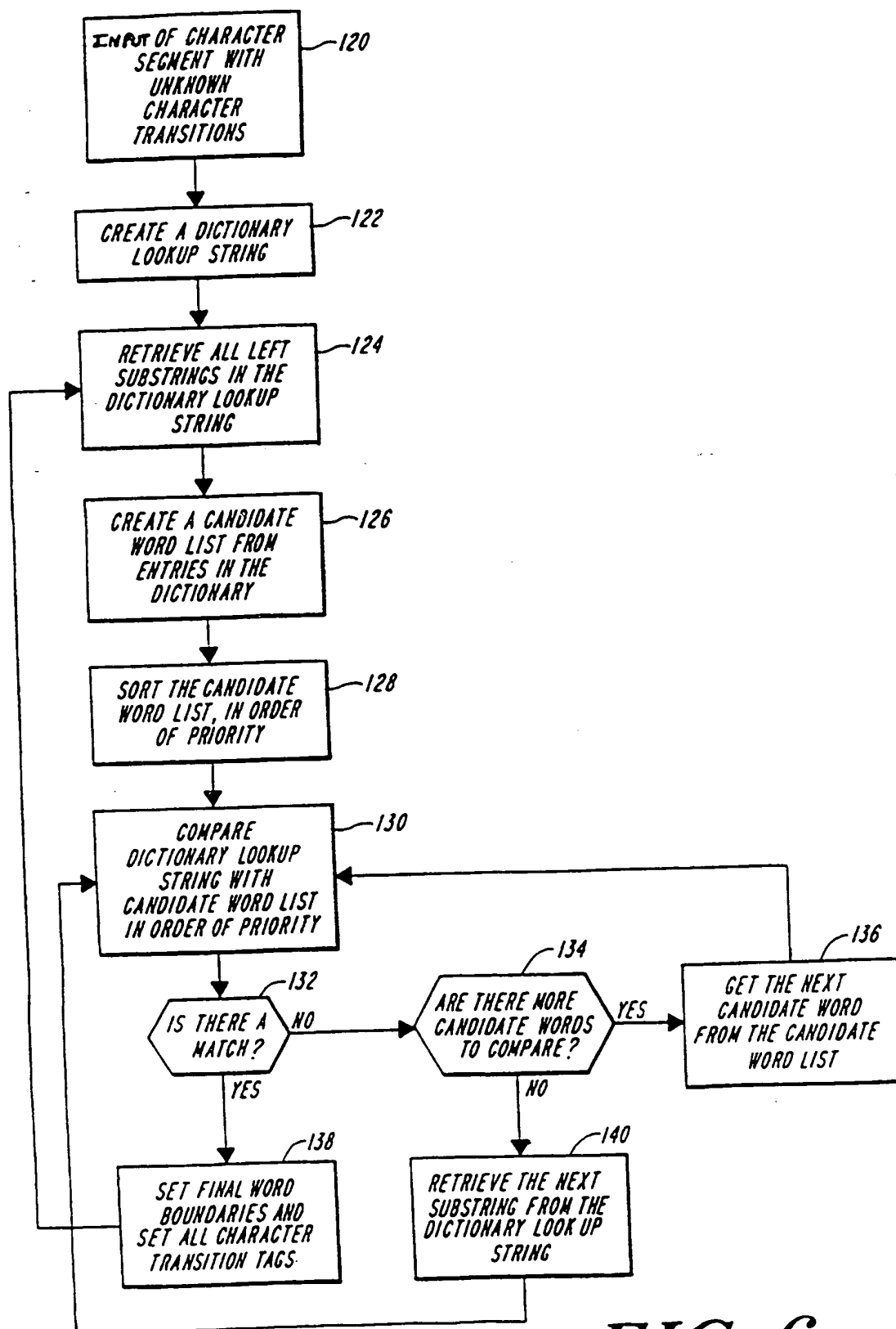


FIG. 6

INTERNATIONAL SEARCH REPORT

International Application No.

PCT/US 97/14741

A. CLASSIFICATION OF SUBJECT MATTER

IPC 6 G06F17/27

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	V. TELLER ET AL: "A probabilistic algorithm for segmenting non-Kanji Japanese strings" PROCEEDINGS OF THE TWELFTH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, SEATTLE, US, vol. 1, 31 July 1994 - 4 August 1994, MIT PRESS, CAMBRIDGE, US, pages 742-747, XP000612334 see page 742, right-hand column, last paragraph - page 743, right-hand column, paragraph 3 see page 744, right-hand column, paragraph 2 - page 745, left-hand column, paragraph 1 --- -/--	1-41



Further documents are listed in the continuation of box C



Patent family members are listed in annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"3" document member of the same patent family

Date of the actual completion of the international search

27 November 1997

Date of mailing of the international search report

04/12/1997

Name and mailing address of the ISA

European Patent Office, P.O. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Pottiez, M

INTERNATIONAL SEARCH REPORT

International Application No.

PCT/US 97/14741

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 029 084 A (MOROHASI ET AL) 2 July 1991 cited in the application see column 2, line 64 - line 68 see column 4, line 42 - line 65 ---	1-41
A	US 5 268 840 A (CHANG ET AL) 7 December 1993 cited in the application see abstract ---	1-41
A	EP 0 547 222 A (CSK K.K.) 23 June 1993 see page 3, line 46 - line 49 see page 7, line 30 - line 48 -----	1,27,40

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 97/14741

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5029084 A	02-07-91	JP 1234975 A	20-09-89
		JP 1807961 C	10-12-93
		JP 5019184 B	16-03-93
US 5268840 A	07-12-93	JP 7311768 A	28-11-95
EP 0547222 A	23-06-93	JP 4211867 A	03-08-92
		CA 2070406 A	21-02-92
		WO 9203796 A	05-03-92

THIS PAGE BLANK (USPTO)